Towards a noisy-channel model of dysarthria in speech recognition

Frank Rudzicz
University of Toronto, Department of Computer Science
Toronto, Ontario, Canada
frank@cs.toronto.edu

Abstract

Modern automatic speech recognition is ineffective at understanding relatively unintelligible speech caused by neuro-motor disabilities collectively called dysarthria. Since dysarthria is primarily an articulatory phenomenon, we are collecting a database of vocal tract measurements during speech of individuals with cerebral palsy. In this paper, we demonstrate that articulatory knowledge can remove ambiguities in the acoustics of dysarthric speakers by reducing entropy relatively by 18.3%, on average. Furthermore, we demonstrate that dysarthric speech is more precisely portrayed as a noisy-channel distortion of an abstract representation of articulatory goals, rather than as a distortion of non-dysarthric speech. We discuss what implications these results have for our ongoing development of speech systems for dysarthric speakers.

1 Introduction

Dysarthria is a set of congenital and traumatic neuro-motor disorders that impair the physical production of speech and affects approximately 0.8% of individuals in North America (Hosom et al., 2003). Causes of dysarthria include cerebral palsy (CP), multiple sclerosis, Parkinson's disease, and amyotrophic lateral sclerosis (ALS). These impairments reduce or remove normal control of the primary vocal articulators but do not affect the abstract production of meaningful, syntactically correct language.

The neurological origins of dysarthria involve damage to the cranial nerves that control the speech articulators (Moore and Dalley, 2005). Spastic dysarthria, for instance, is partially caused by lesions in the facial and hypoglossal nerves, which control the jaw and tongue respectively (Duffy, 2005), resulting in slurred speech and a less differentiable vowel space (Kent and Rosen, 2004). Similarly, damage to the glossopharyngeal nerve can reduce control over vocal fold vibration (i.e., phonation), resulting in guttural or grating raspiness. Inadequate control of the soft palate caused by disruption of the vagus nerve may lead to a disproportionate amount of air released through the nose during speech (i.e., hypernasality).

Unfortunately, traditional automatic speech recognition (ASR) is incompatible with dysarthric speech, often rendering such software inaccessible to those whose neuro-motor disabilities might make other forms of interaction (e.g., keyboards, touch screens) laborious. Traditional representations in ASR such as hidden Markov models (HMMs) trained for speaker independence that achieve 84.8% word-level accuracy for non-dysarthric speakers might achieve less than 4.5% accuracy given severely dysarthric speech on short sentences (Rudzicz, 2007). Our research group is currently developing new ASR models that incorporate empirical knowledge of dysarthric articulation for use in assistive applications (Rudzicz, 2009). Although these models have increased accuracy, the disparity is still high. Our aim is to understand *why* ASR fails for dysarthric speakers by understanding the acoustic and articulatory nature of their speech.

In this paper, we cast the speech-motor interface within the mathematical framework of the noisy-channel model. This is motivated by the charac-

Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies, pages 80–88,
Los Angeles, California, June 2010.

terization of dysarthria as a distortion of parallel biological pathways that corrupt motor signals before execution (Kent and Rosen, 2004; Freund et al., 2005), as in the examples cited above. Within this information-theoretic framework, we aim to infer the nature of the motor signal distortions given appropriate measurements of the vocal tract. That is, we ask the following question: Is dysarthric speech a distortion of typical speech, or are they both distortions of some common underlying representation?

2 Dysarthric articulation data

Since the underlying articulatory dynamics of dysarthric speech are intrinsically responsible for complex acoustic irregularities, we are collecting a database of dysarthric articulation. Time-aligned movement and acoustic data are measured using two systems. The first infers 3D positions of surface facial markers given stereo video images. The second uses electromagnetic articulography (EMA), in which the speaker is placed within a cube that produces a low-amplitude electromagnetic field, as shown in figure 1. Tiny sensors within this field allow the inference of articulator positions and velocities to within 1 mm of error (Yunusova et al., 2009).

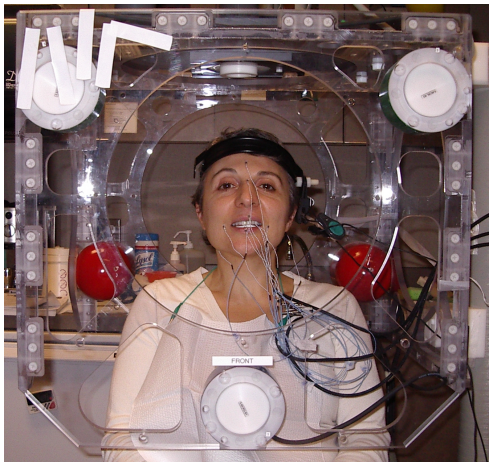


Figure 1: Electromagnetic articulograph system.

We have so far recorded one male speaker with ALS, five male speakers with CP, four female speakers with CP, and age- and gender-matched controls. Measurement coils are placed as in other studies (e.g., the University of Edinburgh's MOCHA database (Wrench, 1999) and the University of Wisconsin-Madison's x-ray microbeam database (Yunusova et al., 2008)). Specifically, we are interested in the positions of the upper and lower lip (UL and LL), left and right mouth corners (LM and RM), lower incisor (LI), and tongue tip, blade, and dorsum (TT, TB, and TD). Unfortunately, a few of our male CP subjects had a severe gag reflex, and we found it impossible to place more than one coil on the tongue for these few individuals. Therefore, of the tongue positions, only TT is used in this study. All articulatory data are smoothed with third-order median filtering in order to minimize measurement 'jitter'. Figure 2 shows the degree of lip aperture (i.e., the distance between UL and LL) over time for a control and a dysarthric speaker repeating the sequence */ah p iy/*. Here, the dysarthric speech is notably slower and has more excessive movement.

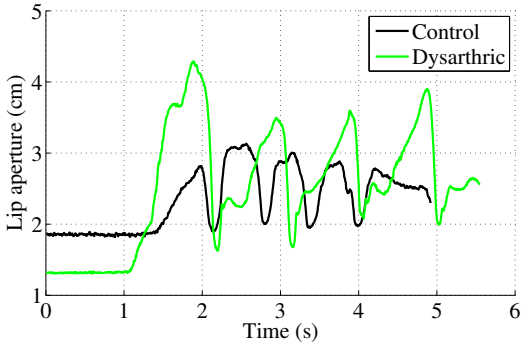


Figure 2: Lip aperture over time for four iterations of */ah p iy/* given a dysarthric and control speaker.

Our dysarthric speech data include random repetitions of phonetically balanced short sentences originally used in the TIMIT database (Zue et al., 1989), as well as pairs of monosyllabic words identified by Kent et al. (1989) as having relevant articulatory contrasts (e.g., *beat* versus *meat* as a stop-nasal contrast). All articulatory data are aligned with associated acoustic data, which are transformed to Mel-frequency cepstral coefficients (MFCCs). Phoneme boundaries and pronunciation errors are being transcribed by a speech-language pathologist to the TIMIT phoneset. Table 1 shows pronunciation errors according to manner of articulation for dysarthric speech. Plosives are mispronounced most often, with substitution errors exclusively caused by errant voicing (e.g. */d/* for */t/*). By comparison, only

5% of corresponding plosives in total are mispronounced in regular speech. Furthermore, the prevalence of deleted affricates in word-final positions, almost all of which are alveolar, does not occur in the corresponding control data.

	SUB (%)			DEL (%)		
	i	m	f	i	m	f
plosives	13.8	18.7	7.1	1.9	1.0	12.1
affricates	0.0	8.3	0.0	0.0	0.0	23.2
fricatives	8.5	3.1	5.3	22.0	5.5	13.2
nasals	0.0	0.0	1.5	0.0	0.0	1.5
glides	0.0	0.7	0.4	11.4	2.5	0.9
vowels	0.9	0.9	0.0	0.0	0.2	0.0

Table 1: Percentage of phoneme substitution (SUB) and deletion (DEL) errors in word-initial (i), word-medial (m), and word-final (f) positions across categories of manner for dysarthric data.

Table 2 shows the relative durations of the five most common vowels and sonorant consonants in our database between dysarthric and control speech. Here, dysarthric speakers are significantly slower than their control counterparts at the 95% confidence interval for */eh/* and at the 99.5% confidence interval for all other phonemes.

Phoneme	duration (μ (σ^2), in ms)		Avg.
	Dysarthric	Control	diff.
/ah/	189.3 (19.2)	120.1 (4.0)	69.2
/ae/	211.6 (16.4)	140.0 (4.4)	71.6
/eh/	160.5 (7.4)	107.3 (2.6)	53.2
/iy/	177.1 (86.7)	105.8 (93.1)	71.3
/er/	220.5 (27.9)	148.6 (59.8)	71.9
/l/	138.5 (8.0)	91.8 (2.4)	46.7
/m/	173.5 (13.4)	94.7 (2.1)	78.8
/n/	168.4 (14.4)	90.9 (2.3)	77.5
/r/	138.8 (8.3)	95.3 (3.4)	43.5
/w/	151.5 (12.0)	84.5 (1.3)	67.0

Table 2: Average lengths (and variances in parentheses) in milliseconds for the five most common vowels and sonorant consonants for dysarthric and control speakers. The last column is the average difference in milliseconds between dysarthric and control subjects.

Processing and annotation of further data from additional dysarthric speakers is ongoing, including measurements of all three tongue positions.

3 Entropy and the noisy-channel model

We wish to measure the degree of statistical disorder in both acoustic and articulatory data for dysarthric and non-dysarthric speakers, as well as the *a posteriori* disorder of one type of data given the other. This quantification will inform us as to the relative merits of incorporating knowledge of articulatory behaviour into ASR systems for dysarthric speakers. Entropy, $H(X)$, is a measure of the degree of uncertainty in a random variable X. When X is discrete, this value is computed with the familiar

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

where b is the logarithm base, x_i is a value of X, of which there are n possible, and $p(x_i)$ is its probability. When our observations are continuous, as they are in our acoustic and articulatory database, we must use *differential entropy* defined by

$$H(X) = -\int_X f(X) \log f(X) dX,$$

where $f(X)$ is the probability density function of X. For a number of distributions $f(X)$, the differential entropy has known forms (Lazo and Rathie, 1978). For example, if $f(X)$ is a multivariate normal,

$$f_X(x_1,...,x_N) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}{(2\pi)^{N/2}|\Sigma|^{1/2}} \quad (1)$$
$$H(X) = \tfrac{1}{2}\ln\left((2\pi e)^N |\Sigma|\right),$$

where μ and Σ are the mean and covariances of the data. However, since we observe that both acoustic and articulatory data follow non-Gaussian distributions, we choose to represent these spaces by mixtures of Gaussians. Huber et al. (2008) have developed an accurate algorithm for estimating differential entropy of Gaussian mixtures based on iteratively merging Gaussians and the approximation

$$\tilde{H}(X) = \sum_{i=1}^{L} \omega_i \left(-\log \omega_i + \tfrac{1}{2}\log((2\pi e)^N |\Sigma_i|\right),$$

where ω_i is the weight of the $i^{th}(1 \leq i \leq L)$ Gaussian and Σ_i is that Gaussian's covariance matrix. This method is used to approximate entropies in the following study, with $L = 32$. Note that while differential entropies *can* be negative and not invariant under

change of variables, other properties of entropy are retained (Huber et al., 2008), such as the chain rule for conditional entropy

$$H(Y|X) = H(Y,X) - H(X),$$

which describes the uncertainty in Y given knowledge of X, and the chain rule for mutual information

$$I(Y;X) = H(X) + H(Y) - H(X,Y),$$

which describes the mutual dependence between X and Y. Here, we quantize entropy with the *nat*, which is the natural logarithmic unit, e (≈ 1.44 bits).

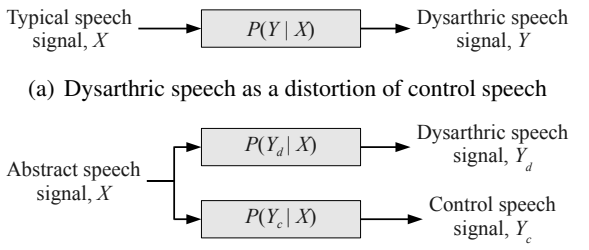
3.1 The noisy channel

The noisy-channel theorem states that information passed through a channel with capacity C at a rate $R \leq C$ can be reliably recovered with an arbitrarily low probability of error given an appropriate coding. Here, a message from a finite alphabet is encoded, producing signal $x \in X$. That signal is then distorted by a medium which transmits signal $y \in Y$ according to some distribution $P(Y|X)$. Given that there is some probability that the received signal, y, is corrupted, the message produced by the decoder may differ from the original (Shannon, 1949).

To what extent can we describe the effects of dysarthria within an information-theoretic noisy channel model? We pursue two competing hypotheses within this general framework. The first hypothesis models the assumption that dysarthric speech is a distorted version of typical speech. Here, signal X and Y represent the vocal characteristics of the general and dysarthric populations, respectively, and $P(Y|X)$ models the distortion between them. The second hypothesis models the assumption that *both* dysarthric and typical speech are distorted versions of some common abstraction. Here, Y_d and Y_c represent the vocal characteristics of dysarthric and control speakers, respectively, and X represents a common, underlying mechanism and that $P(Y_d|X)$ and $P(Y_c|X)$ model distortions from that mechanism. These two hypotheses are visualized in figure 3. In each of these cases, signals can be acoustic, articulatory, or some combination thereof.

3.2 Common underlying abstractions

In order to test our hypothesis that both dysarthric and control speakers share a common high-level abstraction of the vocal tract that is in both cases distorted during articulation, we incorporate the theory of *task dynamics* (Saltzman and Munhall, 1989). This theory represents the interface between the lexical intentions and vocal tract realizations of speech as a sequence of overlapping *gestures*, which are continuous dynamical systems that describe goal-oriented reconfigurations of the vocal tract, such as bilabial closure during */m/*. Figure 4 shows an example of overlapping gestures for the word *pub*.



(a) Dysarthric speech as a distortion of control speech

(b) Dysarthric and control speech as distortions of a common abstraction

Figure 3: Sections of noisy channel models that mimic the neuro-motor interface.

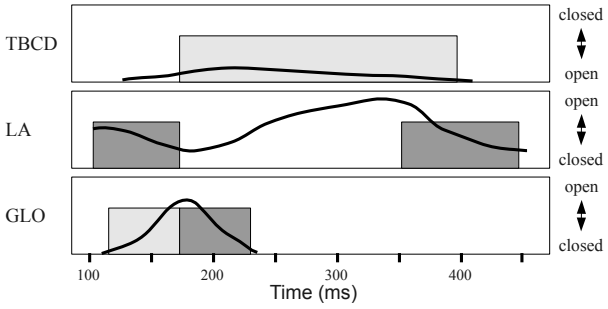


Figure 4: Canonical example *pub* from Saltzman and Munhall (1989) representing overlapping goals for tongue blade constriction degree (TBCD), lip aperture (LA), and glottis (GLO). Boxes represent the present of discretized goals, such as lip closure. Black curves represent the output of the TADA system.

The open-source TADA system (Nam and Goldstein, 2006) estimates the positions of various articulators during speech according to parameters that have been carefully tuned by the authors of TADA according to a generic, speaker-independent representation of the vocal tract (Saltzman and Munhall, 1989). Given a word sequence and a syllable-to-gesture dictionary, TADA produces the continuous

tract variable paths that are necessary to produce that sequence. This takes into account various physiological aspects of human speech production, such as interarticulator co-ordination and timing (Nam and Saltzman, 2003).

In this study, we use TADA to produce estimates of a global, high-level representation of speech common to both dysarthric and non-dysarthric speakers alike. Given a word sequence uttered by both types of speaker, we produce five continuous curves prescribed by that word sequence in order to match our available EMA data. Those curves are lip aperture and protrusion (LA and LP), tongue tip constriction location and degree (TTCL and TTCD, representing front-back and top-down positions of the tongue tip, respectively), and lower incisor height (LIH). These curves are then compared against actually observed EMA data, as described below.

4 Experiments

First, in section 4.1, we ask whether the incorporation of articulatory data is theoretically useful in reducing uncertainty in dysarthric speech. Second, in section 4.2, we ask which of the two noisy channel models in figure 3 best describe the observed behaviour of dysarthric speech.

Data for this study are collected as described as in section 2. Here, we use data from three dysarthric speakers with cerebral palsy (males M01 and M04, and female F03), as well as their age- and gender-matched counterparts from the general population (males MC01 and MC03, and female FC02). For this study we restrict our analysis to 100 phrases uttered in common by all six speakers.

4.1 Entropy

We measure the differential entropy of acoustics ($H(Ac)$), of articulation ($H(Ar)$), and of acoustics given knowledge of the vocal tract ($H(Ac|Ar)$) in order to obtain theoretical estimates as to the utility of articulatory data. Table 3 shows these quantities across the six speakers in this study. As expected, the acoustics of dysarthric speakers are much more disordered than for non-dysarthric speakers. One unexpected finding is that there is very little difference between speakers in terms of their entropy of articulation. Although dysarthric speakers clearly lack articulatory dexterity, this implies that they nonetheless articulate with a level of consistency similar to their non-dysarthric counterparts[1]. However, the equivocation $H(Ac|Ar)$ is an order of magnitude lower for non-dysarthric speakers. This implies that there is very little ambiguity left in the acoustics of non-dysarthric speakers if we have simultaneous knowledge of the vocal tract, but that quite a bit of ambiguity remains for our dysarthric speakers, despite significant reductions.

	Speaker	$H(Ac)$	$H(Ar)$	$H(Ac\|Ar)$
Dys.	M01	66.37	17.16	50.30
	M04	33.36	11.31	26.25
	F03	42.28	19.33	39.47
	Average	47.34	15.93	38.68
Ctrl.	MC01	24.40	21.49	1.14
	MC03	18.63	18.34	3.93
	FC02	16.12	15.97	3.11
	Average	19.72	18.60	2.73

Table 3: Differential entropy, in nats, across dysarthric and control speakers for acoustic *ac* and articulatory *ar* data.

Table 4 shows the average mutual information between acoustics and articulation for each type of speaker, given knowledge of the phonological manner of articulation. In table 1 we noted a prevalence of pronunciation errors among dysarthric speakers for plosives, but table 4 shows no particularly low congruity between acoustics and articulation for this manner of phoneme. Those pronunciation errors tended to be voicing errors, which would involve the glottis, which is not measured in this study.

Table 4 appears to imply that there is little mutual information between acoustics and articulation in vowels across all speakers. However, this is almost certainly the result of our exclusion of tongue blade and tongue dorsum measurements in order to standardize across speakers who could not manage these sensors. Indeed, the configuration of the entire tongue is known to be useful in discriminating among the vowels (O'Shaughnessy, 2000). An *ad hoc* analysis including all three tongue sensors for speakers F03, MC01, MC03, and FC02 revealed mutual information between acoustics and articula-

[1] This is borne out in the literature (Kent and Rosen, 2004).

Manner	$I(Ac;Ar)$	
	Dys.	Ctrl.
plosives	10.92	16.47
affricates	8.71	9.23
fricatives	9.30	10.94
nasals	13.29	15.10
glides	11.92	12.68
vowels	6.76	7.15

Table 4: Mutual information $I(Ac;Ar)$ of acoustics and articulation for dysarthric and control subjects, across phonological manners of articulation.

tion of 16.81 nats for F03 and 18.73 nats for the control speakers, for vowels. This is compared with mutual information of 11.82 nats for F03 and 13.88 nats for the control speakers across all other manners. The trend seems to be that acoustics are better predicted given more tongue measurements.

In order to better understand these results, we compare the distributions of the vowels in acoustic space across dysarthric and non-dysarthric speech. Vowels in acoustic space are characterized by the steady-state positions of the first two formants (F1 and F2) as determined automatically by applying the pre-emphasized Burg algorithm (Press et al., 1992). We fit Gaussians to the first two formants for each of the vowels in our data, as exemplified in figure 5 and compute the entropy within these distributions. Surprisingly, the entropies of these distributions were relatively consistent across dysarthric (34.6 nats) and non-dysarthric (33.3 nats) speech, with some exceptions (e.g., *iy*). However, vowel spaces overlap considerably more in the dysarthric case signifying that, while speakers with CP can be nearly as acoustically consistent as non-dysarthric speakers, their targets in that space are not as discernible. Some research has shown larger variance among dysarthric vowels relative to our findings (Kain et al., 2007). This may partially be due to our use of natural connected speech as data, rather than restrictive consonant-vowel-consonant non-words.

4.2 Noisy channel

Our task is to determine whether dysarthric speech is best represented as a distorted version of typical speech, or if both dysarthric and typical speech ought to be viewed as distortions of a common ab-

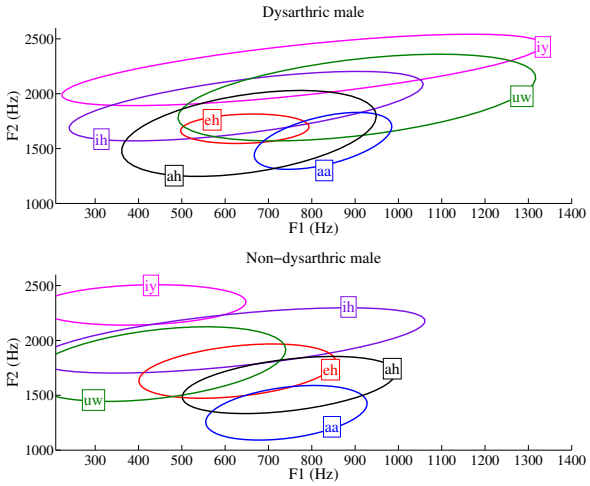


Figure 5: Contours showing first standard deviation in F1 versus F2 space for distributions of six representative vowels in continuous speech for the dysarthric and non-dysarthric male speakers.

stract representation. To explore this question, we design a transformation system that produces the most likely observation in one data space given its counterpart in another and the statistical relationship between the two spaces. This transformation in effect implements the noisy channel itself.

To accomplish this, we learn probability distributions over our EMA data. First, we collect all dysarthric data together and all non-dysarthric data together. We then consider the acoustic (Ac) and articulatory (Ar) subsets of these data. In each case, we train Gaussian mixtures, each with 60 components, over 90% of the data in both dysarthric and non-dysarthric speech. Here, each of the 60 phonemes in the data is represented by one Gaussian component, with the weight of that component determined by the relative proportion of 10 ms frames for that phoneme. Similarly, all training word sequences are passed to TADA, and we train a mixture of Gaussians on its articulatory output.

Across all Gaussian mixtures, we end up with 5 Gaussians tuned to various aspects of each phoneme p: its dysarthric acoustics and articulation ($\mathbf{N}_p^{Ac}(Y_d)$ and $\mathbf{N}_p^{Ar}(Y_d)$), its control acoustics and articulation ($\mathbf{N}_p^{Ac}(Y_d)$ and $\mathbf{N}_p^{Ar}(Y_d)$), and its prescribed articulation from TADA ($\mathbf{N}_p^{Ar}(X)$). Each Gaussian $\mathbf{N}_p^{A}(B)$ is represented by its mean $\mu_p^{(A,B)}$ and its

covariance, $\Sigma_p^{(A,B)}$. Furthermore, we compute the cross-covariance matrix between Gaussians for a given phoneme (e.g., $\Sigma_p^{(Ac,Y_c)\rightarrow(Ac,Y_d)}$ is the cross-covariance matrix of the acoustics of the control (Y_c) and dysarthric (Y_d) speech for phoneme p). Given these parameters, we estimate the most likely frame in one domain given its counterpart in another. For example, if we are given a frame of acoustics from a control speaker, we can synthesize the most likely frame of acoustics for a dysarthric speaker, given an application of the noisy channel proposed by Hosom et al. (2003) used to transform dysarthric speech to make it more intelligible. Namely, given a frame of acoustics y_c from a control speaker, we can estimate the acoustics of a dysarthric speaker y_d with:

$$\begin{aligned} f_{Ac}(y_c) =& E(y_d \,|\, y_c) \\ =& \sum_{i=1}^{P} h_i(y_c) \left[\mu_i^{(Ac,Y_d)} + \right. \\ & \left. \Sigma_i^{(Ac,Y_c)\rightarrow(Ac,Y_d)} \cdot \left(\Sigma_i^{(Ac,Y_c)}\right)^{-1} \cdot \right. \\ & \left. \left(y_c - \mu_i^{(Ac,Y_c)}\right) \right], \end{aligned} \tag{2}$$

where

$$h_i(y_c) = \frac{\alpha_i N\left(y_c; \mu_i^{(Ac,Y_c)}, \Sigma_i^{(Ac,Y_c)}\right)}{\sum_{j=1}^{P} \alpha_j N\left(y_c; \mu_j^{(Ac,Y_c)}, \Sigma_j^{(Ac,Y_c)}\right)},$$

where α_p is the proportion of the frames of phoneme p in the data. Transforming between different types and sources of data is accomplished merely by substituting in the appropriate Gaussians above.

We now measure how closely the transformed data spaces match their true target spaces. In each case, we transform test utterances (recorded, or synthesized with TADA) according to functions learned in training (i.e., we use the remaining 10% of the data for each speaker type). These transformed spaces are then compared against their target space in our data. Table 5 shows the Gaussian mixture phoneme-level Kullback-Leibler divergences given various types of source and target data, weighted by the relative proportions of the phonemes. Each pair of N-dimensional Gaussians ($\mathbf{N}_i$ with mean μ_i and covariance Σ_i) for a given phone and data type is compared with

Type 1	Type 2	KL divergence (10^{-2} nats) Acous.	Artic.
Ctrl.	Dys.	25.36	3.23
Ctrl. → Dys.	Dys.	17.78	2.11
TADA → Ctrl.	Ctrl.	N/A	1.69
TADA → Dys.	Dys.	N/A	1.84

Table 5: Average weighted phoneme-level Kullback-Leibler divergences.

$$\begin{aligned} D_{KL}(\mathbf{N}_0 \,||\, \mathbf{N}_1) =& \frac{1}{2}\left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \mathrm{trace}(\Sigma_1^{-1}\Sigma_0) \right. \\ & \left. + (\mu_1 - \mu_0)^T \Sigma_1^{-1} (\mu_1 - \mu_0) - N \right). \end{aligned}$$

Our baseline shows that control and dysarthric speakers differ far more in their acoustics than in their articulation. When our control data (both acoustic and articulatory) are transformed to match the dysarthric data, the result is predictably more similar to the latter than if the conversion had not taken place. This corresponds to the noisy channel model of figure 3(a), whereby dysarthric speech is modelled as a distortion of non-dysarthric speech. However, when we model dysarthric and control speech as distortions of a common, abstract representation (i.e., task dynamics) as in figure 3(b), the resulting synthesized articulatory spaces are more similar to their respective observed data than the articulation predicted by the first noisy channel model. Dysarthric articulation predicted by transformations from task-dynamics space differ significantly from those predicted by transformations from control EMA data at the 95% confidence interval.

5 Discussion

This paper demonstrates a few acoustic and articulatory features in speakers with cerebral palsy. First, these speakers are likely to mistakenly voice unvoiced plosives, and to delete fricatives regardless of their word position. We suggest that it might be prudent to modify the vocabularies of ASR systems to account for these expected mispronunciations. Second, dysarthric speakers produce sonorants significantly slower than their non-dysarthric counterparts.

This may present an increase in insertion errors in ASR systems (Rosen and Yampolsky, 2000).

Although not quantified in this paper, we detect that a lack of articulatory control can often lead to observable acoustic consequences. For example, our dysarthric data contain considerable involuntary types of velopharyngeal or glottal noise (often associated with respiration), audible swallowing, and stuttering. We intend to work towards methods of explicitly identifying regions of non-speech noise in our ASR systems for dysarthric speakers.

We have considered the amount of statistical disorder (i.e., entropy) in both acoustic and articulatory data in dysarthric and non-dysarthric speakers. The use of articulatory knowledge reduces the degree of this disorder significantly for dysarthric speakers (18.3%, relatively), though far less than for non-dysarthric speakers (86.2%, relatively). In real-world applications we are not likely to have access to measurements of the vocal tract; however, many approaches exist that estimate the configuration of the vocal tract given only acoustic data (Richmond et al., 2003; Toda et al., 2008), often to an average error of less than 1 mm. The generalizability of such work to new speakers (particularly those with dysarthria) without training is an open research question.

We have argued for noisy channel models of the neuro-motor interface assuming that the pathway of motor command to motor activity is a linear sequence of dynamics. The biological reality is much more complicated. In particular, the pathway of verbal motor commands includes several sources of sensory feedback (Seikel et al., 2005) that modulate control parameters during speech (Gracco, 1995). These senses include exteroceptive stimuli (auditory and tactile), and interoceptive stimuli (particularly proprioception and its kinesthetic sense) (Seikel et al., 2005), the disruption of which can lead to a number of production changes. For instance, Abbs et al. (1976) showed that when conduction in the mandibular branches of the trigeminal nerve is blocked, the resulting speech has considerably more pronunciation errors, although is generally intelligible. Barlow (1989) argues that the redundancy of sensory messages provides the necessary input to the motor *planning* stage, which relates abstract goals to motor activity in the cerebellum. As we continue to develop our articulatory ASR models for dysarthric speakers, one potential avenue for future research involves the incorporation of feedback from the current state of the vocal tract to the motor planning phase. This would be similar, in premise, to the DIVA model (Guenther and Perkell, 2004).

In the past, we have shown that ASR systems that adapt non-dysarthric acoustic models to dysarthric data offer improved word-accuracy rates, but with a clear upper bound approximately 75% below the general population (Rudzicz, 2007). Incorporating articulatory knowledge into such adaptation improved accuracy further, but with accuracy still approximately 60% below the general population (Rudzicz, 2009). In this paper, we have demonstrated that dysarthric articulation can be more accurately represented as a distortion of an underlying model of abstract speech goals than as a distortion of non-dysarthric articulation. These results will guide our continued development of speech systems augmented with articulatory knowledge, particularly the incorporation of task dynamics.

Acknowledgments

This research is funded by Bell University Labs, the Natural Sciences and Engineering Research Council of Canada, and the University of Toronto.

References

James H. Abbs, John W. Folkins, and Murali Sivarajan. 1976. Motor Impairment following Blockade of the Infraorbital Nerve: Implications for the Use of Anesthetization Techniques in Speech Research. *Journal of Speech and Hearing Research*, 19(1):19–35.

H.B. Barlow. 1989. Unsupervised learning. *Neural Computation*, 1(3):295–311.

Joseph R Duffy. 2005. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Mosby Inc.

Hans-Joachim Freund, Marc Jeannerod, Mark Hallett, and Ramón Leiguarda. 2005. *Higher-order motor disorders: From neuroanatomy and neurobiology to clinical neurology*. Oxford University Press.

Vincent L. Gracco. 1995. Central and peripheral components in the control of speech movements. In Fredericka Bell-Berti and Lawrence J. Raphael, editors, *Introducing Speech: Contempory Issues, for Katherine Safford Harris*, chapter 12, pages 417–431. American Institute of Physics press.

Frank H. Guenther and Joseph S. Perkell. 2004. A neural model of speech production and its application to studies of the role of auditory feedback in speech. In Ben Maassen, Raymond Kent, Herman Peters, Pascal Van Lieshout, and Wouter Hulstijn, editors, *Speech Motor Control in Normal and Disordered Speech*, chapter 4, pages 29–49. Oxford University Press, Oxford.

John-Paul Hosom, Alexander B. Kain, Taniya Mishra, Jan P. H. van Santen, Melanie Fried-Oken, and Janice Staehely. 2003. Intelligibility of modifications to dysarthric speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, volume 1, pages 924–927, April.

Marco F. Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D. Hanebeck. 2008. On entropy approximation for Gaussian mixture random vectors. In *Proceedings of the 2008 IEEE International Conference on In Multisensor Fusion and Integration for Intelligent Systems*, pages 181–188, Seoul, South Korea.

Alexander B. Kain, John-Paul Hosom, Xiaochuan Niu, Jan P.H. van Santen, Melanie Fried-Oken, and Janice Staehely. 2007. Improving the intelligibility of dysarthric speech. *Speech Communication*, 49(9):743–759, September.

Ray D. Kent and Kristin Rosen. 2004. Motor control perspectives on motor speech disorders. In Ben Maassen, Raymond Kent, Herman Peters, Pascal Van Lieshout, and Wouter Hulstijn, editors, *Speech Motor Control in Normal and Disordered Speech*, chapter 12, pages 285–311. Oxford University Press, Oxford.

Ray D. Kent, Gary Weismer, Jane F. Kent, and John C. Rosenbek. 1989. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54:482–499.

Aida C. G. Verdugo Lazo and Pushpa N. Rathie. 1978. On the entropy of continuous probability distributions. *IEEE Transactions on Information Theory*, 23(1):120–122, January.

Keith L. Moore and Arthur F. Dalley. 2005. *Clinically Oriented Anatomy, Fifth Edition*. Lippincott, Williams and Wilkins.

Hosung Nam and Louis Goldstein. 2006. TADA (TAsk Dynamics Application) manual.

Hosung Nam and Elliot Saltzman. 2003. A competitive, coupled oscillator model of syllable structure. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, pages 2253–2256, Barcelona, Spain.

Douglas O'Shaughnessy. 2000. *Speech Communications – Human and Machine*. IEEE Press, New York, NY, USA.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C: the art of scientific computing*. Cambridge University Press, second edition.

Korin Richmond, Simon King, and Paul Taylor. 2003. Modelling the uncertainty in recovering articulation from acoustics. *Computer Speech and Language*, 17:153–172.

Kristin Rosen and Sasha Yampolsky. 2000. Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative & Alternative Communication*, 16(1):48–60, Jan.

Frank Rudzicz. 2007. Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech. In *Proceedings of the Ninth International ACM SIGACCESS Conference on Computers and Accessibility*, Tempe, AZ, October.

Frank Rudzicz. 2009. Applying discretized articulatory knowledge to dysarthric speech. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP09)*, Taipei, Taiwan, April.

Elliot L. Saltzman and Kevin G. Munhall. 1989. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4):333–382.

J. Anthony Seikel, Douglas W. King, and David G. Drumright, editors. 2005. *Anatomy & Physiology: for Speech, Language, and Hearing*. Thomson Delmar Learning, third edition.

Claude E. Shannon. 1949. *A Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.

Tomoki Toda, Alan W. Black, and Keiichi Tokuda. 2008. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, 50(3):215–227, March.

Alan Wrench. 1999. The MOCHA-TIMIT articulatory database, November.

Yana Yunusova, Gary Weismer, John R. Westbury, and Mary J. Lindstrom. 2008. Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech, Language, and Hearing Research*, 51:596–611, June.

Yana Yunusova, Jordan R. Green, and Antje Mefferd. 2009. Accuracy Assessment for AG500, Electromagnetic Articulograph. *Journal of Speech, Language, and Hearing Research*, 52:547–555, April.

Victor Zue, Stephanie Seneff, and James Glass. 1989. Speech Database Development: TIMIT and Beyond. In *Proceedings of ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases (SIOA-1989)*, volume 2, pages 35–40, Noordwijkerhout, The Netherlands.